Mineração de padrões no gênero textual blog

Maria Lúcia Barbosa, PGIE/UFRGS - malukroeff@yahoo.com.br
Carlos Emilio Padilla Severo, PGIE/UFRGS – emilio.severo@gmail.com
Eliseo Reategui – PGIE/UFRGS – eliseoreategui@gmail.com

Resumo. Este trabalho apresenta um estudo sobre aplicação de ferramentas para mineração de textos, selecionados do gênero textual emergente blog. Neste estudo, pretende-se descobrir padrões nos termos resultantes da mineração, bem como, verificar se os resultados obtidos podem auxiliar no entendimento do conteúdo dos textos com foco educacional. A mineração dos textos foi realizada através das ferramentas Sobek e Tag Clouds.

Palavras-chave. Mineração de padrões, gêneros textuais, blog.

Mining of patterns in the blog textual genre

Abstract. This work shows a study on application of text mining tools in selected texts from the emerging textual genre blog. In this study, we intend to dicovery patterns in the resulting terms, as well as check that the results may aid in understanding the content of the texts with educational focus. The text mining was performed by Sobek tool and Tag Clouds.

Keywords. Mining of patterns, textual genres, blog.

1. Introdução

A internet tem demonstrado um grande potencial para diversificadas manifestações linguísticas, onde surgem novas linguagens com base em tecnologias que formam gêneros textuais emergentes, tais como: sites, blogs, e-mail, mensagens instantâneas, etc.

Atualmente, a Web é a maior fonte de informação eletrônica que dispomos e os blogs são um dos recursos de publicação mais utilizados naquilo que Tim Berners-Lee, considerado o criador da World Wide Web (WWW), chamou de Web da leitura/escrita (W3, 2008). Conforme Marinho (2007), o blog integra o que se chama de software social, uma vez que é definido como ferramenta que potencializa habilidades sociais e colaborativas humanas, como um meio para facilitar conexões sociais e o intercâmbio de informações e, também, como uma ecologia, pois possibilita um sistema de pessoas, práticas, valores e tecnologias em um ambiente particular.

Mineração de dados é a extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir de um conjunto de dados (WHITEN; FRANK, 2005). Na Web, concentra-se em três atividades básicas:

- Mineração de conteúdo: refere-se a minerar conteúdos textuais de páginas da Web;
- Mineração de estrutura: diz respeito à mineração de links entre as páginas da Web;
- Mineração de uso: concentra-se, basicamente, na mineração dos hábitos e ações tomadas pelos usuários da Web (ZAIANE, 2000).

O enfoque deste estudo será na mineração de conteúdo, uma vez que a extração de dados dar-se-á através de dois textos selecionados que abordam as novas tecnologias da educação. Visa-se, também, comparar os resultados analisados através de duas ferramentas de mineração: Sobek e Tag Clouds, buscando informações relevantes e padrões textuais dos blogs analisados, objetivando constatar a eficiência do uso das mesmas para esse tipo de processo dentro do contexto educacional. Dessa forma, este artigo apresenta os resultados da análise do gênero textual blog através da descrição de suas características, bem como a comparação das funcionalidades de duas ferramentas de mineração de textos.

As próximas seções estão divididas da seguinte forma: a seção 2 apresenta o blog como gênero textual, abordando suas características principais; a seção 3 trata sobre a mineração de textos e apresenta as ferramentas utilizadas para esse estudo; a seção 4 descreve a análise dos resultados obtidos; a seção 5 apresenta alguns trabalhos relacionados e a seção 6 traz as considerações finais.

2. Blog como gênero textual

A literatura relata que Jorn Barger foi o idealizador e criador do termo *weblog*, o qual é um acrônimo que referencia o uso da *World Wide Web* para registro diário de conteúdos nas mais diversas áreas e interesses. Os textos de um *blog* são registrados na forma de pequenos blocos e organizados em uma linha de tempo, de acordo com a data de publicação. Os pequenos blocos de texto de um *blog* podem conter muitos elos para acesso externo a fontes (BLOOD, 2002). As primeiras versões atuavam como uma forma de publicação eletrônica, visto que se baseavam na expressão individual de idéias, e não simplesmente como diários eletrônicos. Atualmente, entre os diversos serviços de *blogs* encontrados na *Web*, pode-se observar o uso intensivo de uma importante ferramenta: o comentário. O comentário é um espaço importante, possibilitando a comunicação e interação entre os participantes. A ferramenta comentário permite que os visitantes incluam observações e contribuições às postagens do autor do *blog* (MARCUSHI; XAVIER, 2005).

Devido aos avanços das tecnologias digitais, os *blogs* incorporam vários recursos, e além de arquivos de textos, também há a possibilidade de registrar idéias, pensamentos e relatos através de arquivos de imagens e sons. Há *sites* que hospedam recursos que permitem a convergência de mídias (texto, imagens e sons em um só lugar), como é o caso do **multiply.com** (MARINHO, 2007).

Hoje em dia, qualquer pessoa, com um pouco de conhecimento do uso do computador e tendo acesso à internet, pode criar um blog, uma vez que a tecnologia empregada para criação e publicação é simples. Também deve ser alguém que goste de

ler, escrever e aprender, tendo tempo para essas atividades. Porém, o *blog* tem algumas características básicas e deve:

- Ser datado;
- Conter seção para apresentação, oportunizando um primeiro contato e entendimento do assunto que é abordado;
- Ser atualizado, na maioria das vezes, diariamente;
- Oportunizar a utilização de textos, cores e imagens, além de vídeo e áudio;
- Apresentar seção para divulgação;
- Apresentar seção de notícias, eventos, artigos de opinião, etc;
- Apresentar seção de comentários;
- Verificar se os links postados funcionam corretamente;
- Geralmente, linguagem informal e espontânea;
- Geralmente, uso de verbo na primeira pessoa, no presente ou passado,
- Não ser privativo.

Além disso, apresenta alguns elementos fundamentais, tais como: corpo (*body*); cabeçalho onde costuma ficar o título do *blog* (*header*); a área total da postagem e das colunas (*outer-wrapper*); coluna de *posts* (*main-wrapper*); coluna propriamente dita (*sidebar-wrapper*); nova coluna (*newsidebar*); e o rodapé (*footer*). Considerando-se alguns parâmetros desse gênero emergente, destacam-se:

1. **Tempo**: predominantemente assíncrono;
2. **Participantes**: bilateral ou multilateral;
3. **Relação entre os participantes:** possibilita a relação entre participantes, tanto do "dono" ou "donos" do blog, quanto dos participantes (amizade, trabalho, notícias, etc).
4. **Tema:** diversos (jornalístico, pessoal, temático, etc).
5. **Recuperação de mensagens:** depende do domínio do dono do blog.
6. **Estilo:** depende do tema
7. **Função:** disponibilizar um assunto de interesse na web, aumentando as possibilidades de acesso às informações referente ao tema selecionado, além de promover a interação social e o pensamento analítico e crítico dos participantes.
8. **Método de armazenamento, busca e gerenciamento de textos:** dependente do domínio do dono do blog, que estipula tempo de armazenamento, e gerencia ou media textos e comentários, por exemplo.
9. **Tipo de conteúdo veiculado**: texto (principal), além de possibilitar textos complementares, dicas de links, imagens, vídeos e áudios, bem como notícias atualizadas relacionadas ao assunto, a fim de movimentar os debates no blog.

10. **Tamanho do conteúdo:** dependente dos limites impostos pelo servidor de serviços de blog.

O *blog* é um legítimo representante de gêneros do discurso digital, o qual possui contextos e aplicações diversas, abrangendo diversas áreas como educação, jornalismo, psicologia, informática, pedagogia, etc. Segundo Maruschi (2005), o blog pode ser considerado um gênero digital, assim como chat, fóruns e e-mails. Por outro lado, alguns autores não consideram blog como um gênero específico, mas somente um espaço de comunicação onde textos e fotos são combinados para a transmissão de idéias, onde o autor não necessita de conhecimentos técnicos (SCHITTINE, 2004).

Considerando-se uma visão educacional para o uso do *blog*, Richardson (2006) aponta como uma ferramenta "construtivista de aprendizagem", onde o que é produzido de relevante pelos alunos e professores vai além da sala de aula. Como exercício da escrita, possibilita o efetivo exercício das etapas que a caracterizam, como rascunho, edição, organização, pré-escrita, leitura da prova, publicação e revisão, uma vez que os alunos podem e devem, primeiramente, produzir um rascunho dos seus *posts*. Com tudo isso, é aceitável avaliar o *blog* como um importante instrumento de escrita colaborativa, justamente porque as mensagens a ele associadas podem ser vistas por outros leitores, e esses podem acrescentar algumas informações na forma de comentários.

3. Mineração de Textos

A mineração de textos apresenta objetivo semelhante à mineração de dados, pois visa encontrar informações relevantes nos documentos analisados através da identificação de padrões. A principal diferença entre os métodos se encontra no fato de que a mineração de dados lida com dados estruturados, e a mineração de textos lida com conteúdo não estruturado, ou semi-estruturado. A mineração de dados pode ser usada para descobrir conhecimento, gerar perfis e análises da evolução da ciência, de técnicas, tecnologias, patentes e recursos humanos, da internet e monitoramento do macro ambiente, em particular, o monitoramento de um tema, um produto, um político ou uma empresa na mídia, de concorrentes ou de todo um setor da economia (FILHO, 2008). Ou seja, é uma busca por padrões.

As ferramentas utilizadas para minerar acabam extraindo centenas de características dos textos, deixando o usuário com muita informação sem utilidade. Devido a isso, muitas técnicas são aplicadas para "enxugar" os resultados oriundos desse processo, entre elas, a busca por tendências feita através da análise de vários documentos, e não somente de um. Algoritmos são aplicados para encontrar padrões e algumas conexões, e o resultado é apresentado através de uma ferramenta de visualização de dados.

Uma das formas de encontrar padrões em textos é através da bibliometria de co-ocorrência de palavras: se duas palavras aparecem juntas no mesmo documento, podem estar conectadas, ou ainda, se as mesmas duas palavras aparecem juntas em muitos documentos, há uma relação entre elas (PORTER, 2006). O processo de mineração auxilia no entendimento e mapeamento da questão a ser respondida, através da identificação de bases de dados adequadas, busca, recuperação, limpeza, análise, interpretação e representação da informação na forma mais de visualização mais efetiva.

Segundo Borges e Domingues (2008), o software Sobek funciona de duas maneiras. Primeiramente, é necessário copiar e colar um texto na interface de entrada de dados e selecionar a opção "minerar texto". A procura por palavras registra ocorrências de palavras repetidas ou sinônimas no documento, fazendo relações e criando grafos de interação entre elas, expondo os principais termos/conceitos do texto em mineração. A partir daí, a ferramenta cria um banco de conceitos e possíveis relações associativas para ajudar na procura de palavras-chave para a funcionalidade da vida simulada das formas no texto selecionado e, como resposta, o software tentará encontrar um maior número de ligações entre os conceitos.

Tag Clouds (2008), ou nuvem de tags, geralmente reúne um conjunto de tags utilizadas em um determinado website disposto em ordem alfabética, e o volume de conteúdos que o site apresenta em cada tag é mostrado proporcionalmente pelo tamanho da fonte. O site www.tagcrowd.com, utilizado neste trabalho para mineração através de tag clouds, proporciona a "criação de sua própria nuvem de tags", a partir de uma seleção de qualquer texto de interesse do usuário para visualização da freqüência de palavras. As opções que esta ferramenta proporciona é quanto à linguagem do texto (idioma), seleção do número máximo de palavras para apresentação, mínimo de frequência com que essas palavras devem aparecer no texto para serem selecionadas, entre outras que não foram utilizadas neste trabalho, como palavras similares apenas para língua inglesa.

4. Estudo Comparativo entre Sobek e Tag Clouds

Para a elaboração do estudo comparativo entre duas ferramentas de extração de termos de um gênero textual, foram utilizados dois textos sobre o uso de tecnologias na educação. Tanto os títulos, quanto os autores de cada texto serão omitidos neste artigo.

Porém, importante salientar que o Texto 1 destaca a utilização de blogs em uma monografia sobre a interatividade na aprendizagem colaborativa. São apresentadas as dificuldades encontradas no desenvolvimento do trabalho, bem como alguns comentários sobre o uso de novas tecnologias na escola e o fomento ao debate e troca de experiências com tecnologias no apoio à educação. O Texto 2 apresenta comentários sobre um relatório técnico proveniente de uma pesquisa acerca da utilização de novas tecnologias por professores e alunos na escola.

Na aplicação de ambas as ferramentas, Sobek e Tag Clouds, optou-se por textos na língua portuguesa em um primeiro experimento. Além disso, nas configurações iniciais das ferramentas, definiu-se a apresentação de no máximo vinte termos resultantes no processo de mineração, onde a frequência mínima dos termos no texto deveria ser de duas ocorrências. As duas ferramentas apresentam a frequência ocorrida ao lado de cada termo.

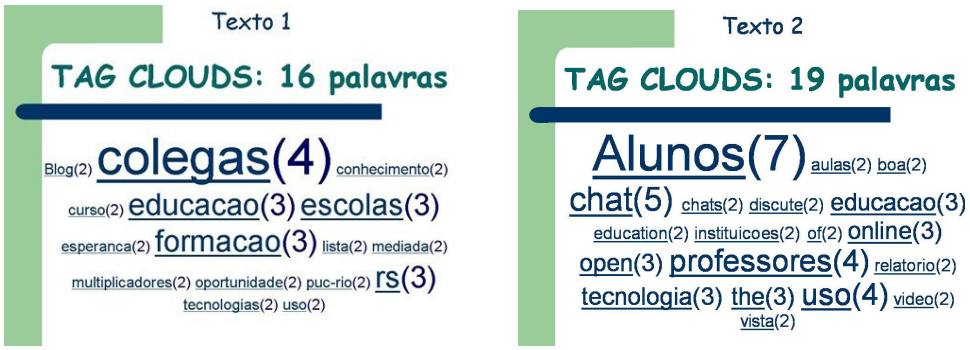
As próximas seções apresentam os resultados obtidos pela aplicação das ferramentas (Tag Clouds e Sobek) para extração de termos das duas amostras de textos.

4.1. Resultados de Aplicação de Tag Clouds

A aplicação da técnica de Tag Clouds no primeiro texto resultou a extração dos termos constantes que aparecem na Figura 1a. Pode-se observar o número de termos

encontrados através da aplicação da técnica de Tag Clouds (16 palavras), bem como, a lista de palavras encontradas com suas respectivas frequências.

A aplicação da mesma técnica no segundo texto resultou a extração dos termos constantes e apresenta o número de termos encontrados através da aplicação da técnica de Tag Clouds, além do rol de palavras com suas frequências (Figura 1b).



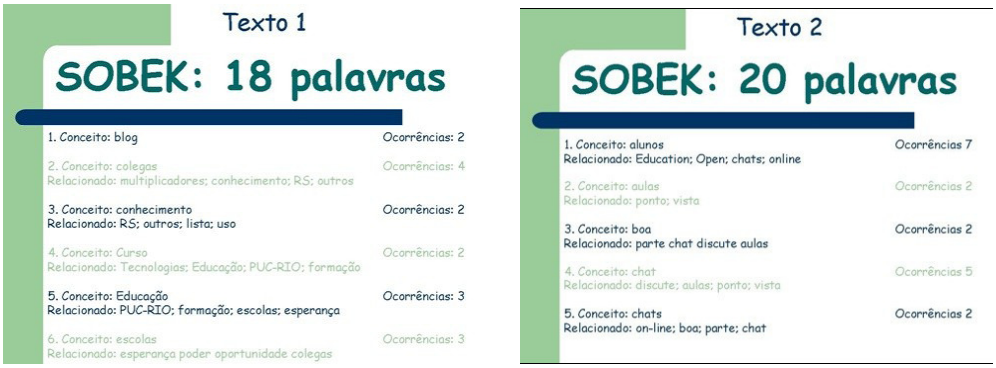
(a) Texto 1 **(b) Texto 2**

Figura 1. Resultado da aplicação de Tag Clouds

4.2. Resultados de Aplicação da Ferramenta Sobek

A ferramenta Sobek apresenta uma lista de termos extraídos de um texto, assim como suas frequências, na mesma forma que a ferramenta utilizada para aplicação da técnica de Tag Clouds. Entretanto, Sobek gera uma lista de relações entre os termos extraídos de um texto. Estas relações são obtidas por uma análise da distância entre os termos de um texto. Quanto mais próximos dois termos estiverem em um texto, mais relacionados estarão.

A Figura 2a apresenta os resultados obtidos pela aplicação da ferramenta Sobek no primeiro texto analisado. Observa-se que o resultado obtido possui duas palavras a mais que por Tag Clouds para o mesmo texto (Figura 1a), além disso, apresenta a lista de palavras encontradas com o número de ocorrências de cada, complementada pela lista de termos relacionados. Na Figura 2a, é mostrada parte da lista obtida na aplicação do Sobek.



(a) Texto 1 **(b) Texto 2**

Figura 2. Resultado da aplicação de Sobek

Observando-se a Figura 2b, nota-se que a ferramenta obteve uma lista de vinte termos no segundo texto analisado, enquanto que para Tag Clouds o resultado foi de 19 palavras para o mesmo texto (Figura 1b). Além das informações textuais obtidas a partir da análise dos dois textos no Sobek, esta ferramenta também gera um grafo apresentando os termos obtidos e seus relacionamentos com demais termos do texto, conforme podemos observar através da Figura 3.

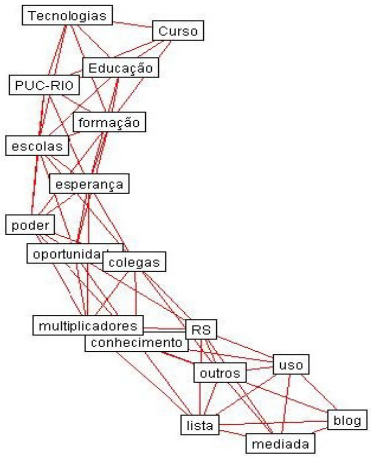


Figura 3. Grafo de relações entre termos do texto 1.

Observam-se, através das Figuras 1 e 2, alguns resultados interessantes comparando-se as diferentes ferramentas de mineração, que podem ser melhor visualizados na Tabela 1 e que é referente ao Texto 2.

Tabela 1. Frequência de palavras do Texto 2 para Tag Clouds e para Sobek

TEXTO 2 / Palavras	TAG CLOUDS	SOBEK
1. Alunos	7	7
2. Aula	2	2
3. Boa	2	2
4. Chat	5	5
5. Chats	2	2
6. Discute	2	2
7. Educação	3	3
8. Education	2	2
9. Instituições	2	2
10. Of	2	não consta
11. Online	3	3
12. Open	3	3
13. Parte	não consta	3
14. Ponto	não consta	2
15. Ponto de Vista	não consta	2
16. Professores	4	4
17. Relatório	2	2

18. Tecnologia	3	3
20. The	3	não consta
21. Uso	4	4
22. Uso do chat	não consta	2
23. Video	2	não consta
24. Vista	2	2

Nota-se que quatro termos que constam na ferramenta Sobek e aparecem de duas a três vezes não foram detectados por Tag Clouds, enquanto três palavras que não foram mineradas por Sobek, aparecem em Tag Clouds, sendo que os requisitos iniciais para mineração foram os mesmos em ambas as ferramentas. E, apesar de um desses requisitos ser "textos em língua portuguesa", foram mineradas palavras em inglês, mostrando falhas nas duas ferramentas.

5. Trabalhos Relacionados

Morais (2007) apresenta um sistema que utiliza técnicas de mineração de textos para associar semanticamente documentos a domínios representados por ontologias. A metodologia proposta, a partir da análise de documentos, usa técnicas estatísticas de mineração de textos, atribuindo um grau de similaridade (ou relevância) desse documento ao domínio representado pela ontologia. No estudo, a ferramenta *Protegé* foi utilizada e testada, e o dados a serem minerados eram documentos contendo jurisprudências do Tribunal de Justiça de Goiás. Assim como no estudo deste artigo, Morais descreve que mesmo existindo várias ferramentas que auxiliam na busca de informações, as mesmas ainda apresentam problemas, principalmente pela dificuldade que os sistemas têm de entender a semântica contida nas páginas, sendo que uma proposta para amenizar tais problemas, é realizar uma análise automática do contexto em que os termos das buscas são usados, permitindo uma "compreensão"do conteúdo, reduzindo ambiguidades e aumentando a relevância dos resultados.

Barbosa (2005) desenvolveu e apresentou a ferramenta MINEGRAF, com o objetivo de facilitar os testes e analisar a performance entre diferentes algoritmos, projetados para diversas tarefas de mineração de dados. A ferramenta é constituída de três módulos: o primeiro permite o cadastro de novos algoritmos, tarefas e parâmetros de mineração, o segundo permite a configuração dos dados de entrada para os algoritmos e o terceiro permite a especificação e visualização de diversos gráficos comparativos entre algoritmos associados à mesma tarefa. Apesar de não ser considerada uma ferramenta de mineração propriamente dita, MINEGRAF provê uma interface gráfica e é indicada para um usuário especialista que esteja interessado em comparar performances de diferentes algoritmos de mineração, pois automatiza a confecção dos mais diversos gráficos de estudo de performance, especificados pelo usuário.

6. Considerações Finais

Uma das contribuições deste artigo, baseando-se em um contexto educacional, foi que dentre essas duas ferramentas, Tag Clouds e Sobek, a segunda tem uma maior

capacidade de extrair informações relevantes, uma vez que relaciona conceitos e apresenta resultados mais completos, além de ser de fácil manejo.

Pode-se perceber ao final da análise dos textos e comparando-se os resultados obtidos pela aplicação das duas ferramentas em dois textos distintos, que a ferramenta Sobek possui características distintas da ferramenta de Tag Clouds, visto que a primeira apresenta uma extração de conceitos encontrados em um texto, acrescidos estes conceitos de relações com outros conceitos encontrados no texto. Já a segunda ferramenta (Tag Clouds), simplesmente realiza uma extração de frequência de termos em um texto. A mesma, em seus resultados, também não apresenta alguns termos encontrados pelo Sobek no primeiro texto analisado, mesmo sendo configurado a frequência mínima para ocorrência de um termo no texto. A análise do segundo texto foi inconsistente entre as duas ferramentas, visto que as palavras encontradas nas duas ferramentas não são equivalentes totalmente.

Com isso, conclui-se que ainda não há uma ferramenta completa, e o ideal é que o usuário leia sobre as características de cada uma e opte por escolher aquela que está mais de acordo com os seus objetivos ou necessidades.

Referências

BARBOSA, F. R.S. MINEGRAF: uma Ferramenta para Testes e Análise Comparativa de Algoritmos de Mineração de Dados. *In: Revista Horizonte Científico*, Uberlândia, MG. 2005.

BLOOD, Rebecca. The Weblog Handbook. Cambridge, MA: Perseus Publishing, 2002.

BORGES, Maicon Mesquita; DOMINGUES, Diana. Mineração de textos para aplicação de algoritmos de vida artificial. XVI Encontro de Jovens Pesquisadores da UCS. Universidade de Caxias do Sul. Caxias do Sul, 2008.

FELDMAN, R.; SANGER, J. Text Mining Handbook. Inglaterra: Universidade de Cambridge, 2006.

MARCUSHI, L. A; XAVIER, A. C. Hipertexto e gêneros digitais: novas formas de construção de sentido. Rio de Janeiro: Lucena, 2005.

MARINHO. Blog na Educação & Manual Básico do Blogger. Programa de Pós-graduação em Educação. PUC Minas, 2007.

MORAIS, E. A. M. Contextualização de Documentos em Domínios Representados por Ontologias Utilizando Mineração de Textos. Dissertação de Mestrado do Instituto de Informática da Universidade Federal de Goiás. Goiás, GO, 2007.

FILHO, R. C. P. Ferramentas para Análise e Mineração de Dados e Textos. In: 2° Seminário sobre Informação na Internet. Conteúdos e Infodiversidade. Empraba, 2008

PORTER, Alan L. VantagePoint Training: Discovering Knowledge on S, T&I Text & Numeric Databases. In: II Seminário Internacional de Ferramentas de Inteligência Competitiva, Brasília - DF, 2006

RICHARDSON, Will. Blogs, wikis, podcasts and other powerful web tools for classroom. Tousand Oaks, USA: Corwin, 2006.

SCHITTINE, Denise. Blog: comunicação e escrita íntima na Internet. Rio de Janeiro: Civilização Brasileira, 2004.

TAG CLOUDS. http://www.tagcrowd.com/. Acesso em dezembro de 2008.

ZAIANE, Osmar R., WEB Mining: Concepts, Practices and Research. In: Simpósio Brasileiro de Banco de Dados, Tutorial, XV SBBD, 2000, João Pessoa. Anais SBBD, João Pessoa: SBBD, 2000. p. 410-474.

WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005

W3. http://www.w3.org/People/Berners-Lee/. Acesso em dezembro de 2008.